A Danish Nonsense Syllable Speech Material

Thomas U. Christiansen, *Centre for Applied Hearing Research, Hearing Systems, Department of Electrical Engineering, Technical University of Denmark, Ørsteds Plads 1, Building 352, DK-2800 Kgs. Lyngby, Denmark*

Email:tuc@elektro.dtu.dk

Abstract

Nonsense syllable speech materials are often used when investigating speech perception in quiet and under adverse conditions. The main advantage of using nonsense syllables over words and sentences is that the acoustic as well as linguistic context is minimal. This paper describes the considerations involved in producing three anechoic recordings of 14 male and 14 female native talkers of Danish each speaking 65 nonsense syllables repeated three times with falling F0 (total of 16380 syllables).

Motivation

Perception of spoken language is a complex process involving several processing stages of quite disparate nature. Such processing stages relate to hearing, lexical structure, phonetic-, morphologic, syntactic and semantic organisation of language. Rather than investigating the process as a whole, we focus on quantitatively characterising the capacity of hearing to identify phonetic segments. Speech material which reduces or eliminates influences of confounding factors, such as processing of lexical structure and morphology, is a requirement. Nonsense syllables are widely used as speech material meeting this requirement as well as possible. In addition, nonsense syllables have minimal acoustic context, which minimises any confounding co-articulatory effects.

The purpose of the present paper is to define a Danish speech material suitable for investigating the role of hearing in phonetic segment identification, and to devise a way to produce such speech material.

Material in English and German

Appendix A shows a small sample of English and German speech material that has been used in the literature. It appears that little effort has gone into describing the material in terms of recording conditions and phonetic design. Moreover, the majority of the speech materials have been designed with a single study in mind.

Basic Phonetic Design of target- and context syllables

The shortest naturally occurring speech sound is arguably (or by definition!) the syllable. The two shortest syllable types are consonant-vowel (CV) and vowel-consonant (VC). Since the former is more frequent in Danish than the latter, the CV syllable type is selected.

The Danish consonants included in this study are shown in Appendix B. They represent phonetic segments with consonantal properties, rather than an exhaustive set based on phonological theory.

Some CV's are also Danish words. It is well-known that it easier to recognise words than non-words. Adding a second syllable could eliminate this perceptual bias. In the remainder of this text the first syllable is referred to as "the target syllable" and the second is referred to as "the context syllable".

Danish has short and long vowels, which may or may not have stød, and which may occur in stressed or unstressed syllables. For simplicity we choose the target syllable to be stressed and to have CVs with long vowels without stød. Ideally all Danish vowels should be used in the target

syllable. For time reasons only [izu] (throughout this paper square brackets designate phones in ttt-notation (Henrichsen, 2007)) were selected for the material presented in this paper. They represent the extremes of the cardinal vowels.

For simplicity the context syllable should be identical for all stimuli. Selecting [tu] as the context syllable entirely eliminates the word bias in that no Danish words exist with the combination of target and context syllables (the latin word [situ] is not a Danish word). Using [t] in the context syllable tends to neutralise co-articulation. Moreover, [t] has a rather central place of articulation, which again is relatively neutral in terms of co-articualtion. In addition, using [t] keeps the lip-rounding from the [u] in the context syllable from spreading to the target syllable.

The disadvantage of selecting [tu] as context syllable is that it is somewhat artificial and that the relative long context "tu" may provide the listener with idiosyncratic cues i.e. cues originating from differences between utterances of the ***context*** (tu) rather than the target CV. A straightforward way of limiting this effect would be to record several utterances for each speaker of the same CV.

Selection of talkers

It is a requirement that the talkers are native speakers of Danish. The talkers will not be formally screened with respect to dialect, voice quality, age, social status etc. Instead each talker will be assessed post hoc, and the assessment will be included in the documentation.

The criteria used for post-hoc assessment of the talkers are:

1. Dialect
2. F0 range
3. etc

Additional Experimental Design

The aim of the experimental design is to ensure that the recorded speech material is as uniform as possible with respect to speed, intonation and vocal effort, and that the target syllables are pronounced in accordance with intended CVs. This can be achieved in a relatively simple procedure whereby a pre-recorded version of the material is played back to the talkers. The talkers are then instructed to repeat the pre-recorded material as faithfully as possible, i.e., same speed, intonation, vocal effort etc.

A *prompting sentence* like "Nu bliver der sagt" preceding the nonsense syllables would serve multiple purposes. 1) It would alert the talker, allowing him or her to focus attention on the nonsense syllable 2) It would reduce any utterance initial problems with voice quality (e.g. irregular vocal fold vibration). 3) The recorded prompting sentence can be used as a prompting sentence in the final speech material.

The pre-recorded nonsense syllables should be repeated three times with falling intonation. There are several reasons for this: 1) It is natural for the readers (fewer mistakes and less variability in the quality of the recorded material). 2) The material will include three instances of the same CV at almost no extra cost in terms of time. 3) The material may be interesting for investigating intonational issues.

In order to assess with-in talker variation, selected CV's will occur multiple times in the recordings. The somewhat arbitrary choice of repeating [vitu] and [vutu] each four additional times was made. The 'v' was chosen because it was speculated that it has a high degree of variability.

In order to improve talker alertness 6 filler syllables with the shown in Appendix C was introduced. The 6 filler sentences have a different context syllable as compared to the proper syllables. The intention is to break any pronunciation adaptation caused by repeating the context syllable.

In total the nonsense syllables recorded will be 17x3 (ordinary target syllables) + 4x2 (repetition syllables [vitu] and [vutu]) + 6 (filler syllables) = 65 syllables. These will be distributed on 5 lists with each 11 syllables and one with 10 syllables).

The simplest playback of the prompting material is to play it back over loudspeakers from a CD player. Playback over headphones would potentially trigger the Lombard effect (Lane H, Tranel B, 1971) and thus illicit a vocal effort deviating from the desired vocal effort. Compensating for this via voice feedback would lead to an overly complicated experimental setup. In order to increase the likelihood of achieving a controlled vocal effort from the talkers, the vocal effort of the talker of the pre-recorded material should be normal (i.e. 65 dB according to ANSI (1997)). Further, the playback level of the pre-recorded material should be set so as to achieve the same level (65 dB).

The CVs should be randomised on the CD (same for different talkers). This is the easiest for subsequent processing of the material. One could consider randomising differently for each talker, but we presuppose this does not affect the spoken material in any significant way. The number of male and female talkers should be reasonably balanced.

Ideally, the pre-recording material should be recorded with both a female and male voice. So female talkers could listen to the pre-recorded material by female talker and male talkers could listen to male version of the pre-recorded material. For time reasons only a male talker was used in the pre-recording.

Two channel recording from head-mounted and table stand microphones will be used. The recordings will take place in the small anechoic chamber at DTU, building 354.

Appendix A: International CV speech materials

ID	Lang-uage	# Speakers	Speech material	Recording	Miscellaneous info	Date	Author	Used by (not full list)
UCLA Nonsense Syllable Test	English	2 females 2 males	CV,18C v=[iOu] O as in bought 8 repititions for each CV. Total of 18x3x8x4=1728 tokens	fs=16kHz	www.ee.ucla.edu/~spapl/cv	1975	Resnick et. al. 1975 ASA Talk	Apoux Bacon 2004 Müsch Buus 2001
LDC-2005S22	US-English	10 female 10 male	CVC VC CV 16 C, V=[u]	fs=16kHz	www.ldc.upenn.edu	2005	J.Wright	Allen 2009
Diagnostic Rhyme Test	US-English	1 male??	Words (minimal pairs)		Designed to test binary phonetic features	1983	Voiers	Voiers, 1983
OLLO	German	40	CVC VCV 10C, 7V	Headset + condenser mic fs=44.1kHz	Four dialects Six speaking styles For HSR-ASR comparisons	2005?	Wesker et. al.	
Consonant Challenge Interspeech	English	12 female 12 male	VCV 24 C V=[iau]	HP – 50Hz fs=25kHz	www.odettes.dds.nl /challenge_IS0 For ASR-ASR comparisons	2008	Scharenborg + Cooke	
Lippmann et. al 1981	English	1 female 1 male	CVC 17 C, 6V	Sound insulated booth.	Reverberation added via playback over loudspeakers		Lippmann et. al. 1981	Lippmann et. al. 1981

Appendix B: List of Danish consonantal sounds considered

The Danish consonants in ttt-notation (a special version of SAMPA representation described in (Henrichsen, 2007)) considered in this study are:

[ptkbdgfsvmnrlhjSW]

Appendix C: Fillers

The six fillers

Sz:ta	mi:ta	ju:ta
lz:ta	ru:ta	Wi:ta

Appendix D: Target CV's

Proposed target CV plus context syllable

pz:tu	pi:tu	pu:tu
tz:u	ti:tu	tu:tu
kz:tu	ki:tu	ku:tu
bz:tu	bi:tu	bu:tu
dz:tu	di:tu	du:tu
gz:tu	gi:tu	gu:tu
fz:tu	fi:tu	fu:tu
sz:tu	si:tu	su:tu
vz:tu	vi:tu	vu:tu
mz:tu	mi:tu	mu:tu
nz:tu	ni:tu	nu:tu
rz:tu	ri:tu	ru:tu
lz:tu	li:tu	lu:tu
hz:tu	hi:tu	hu:tu
jz:tu.	ji:tu.	ju:tu.
Sz:tu	Si:tu	Su:tu
Wz:tu	Wi:tu	Wu:tu

Appendix E: Test lists

The six lists with randomised nonsense syllables, fillers and repeated nonsense syllables.

List 1	List 2	List 3	List 4	List 5	List 6
pz:tu	pi:tu	pu:tu	kz:tu	ki:tu	ku:tu
ru:tu	nu:tu	mi:tu	mz:tu	nz:tu	ni:tu
vi:tu	vz:tu	li:tu	vu:tu	mu:tu	lz:tu
Sz:ta	ju:ta	ru:ta	Wi:ta	lz:ta	mi:ta
ti:tu	tz:tu	dz:tu	bu:tu	tu:tu	bz:tu
vu:tu	fi:tu	fu:tu	fz:tu	si:tu	sz:tu
ha:tu	rz:tu	ri:tu	lu:tu	hu:tu	hi:tu
vi:tu	vi:tu	vu:tu	vu:tu	vi:tu	vu:tu
Wz:tu	Su:tu	Wi:tu	Si:tu	jz:tu	gz:tu
su:tu	ji:tu	Sz:tu	vi:tu	Wu:tu	ju:tu
bi:tu	du:tu	gi:tu	gu:tu	di:tu	

Appendix F: Oral instructions to the talkers

Denne session består af to dele: Gentagelse af nonsens-stavelser samt oplæsning af danske sætninger.

1. Inden den egentlige optagelse går i gang: Spørg om alder, hvor taleren er vokset op og forældrenes talesprog (ganske kort)
2. Lyt og gentag seks lister med 11 nonsens-stavelser i hver
3. Oplæsning af 10 sætninger.
4. Hver sætning læses tre gange umiddelbart efter hinanden
5. Kort vandpause
6. Lyt og gentag seks lister med 11 nonsens-stavelser i hver (samme lister som før)
7. Oplæsning af 10 sætninger. Hver sætning læses tre gange umiddelbart efter hinanden (andre sætninger end tidligere).
8. Lyt og gentag seks lister med 11 nonsens-stavelser i hver (samme lister som før)

Bemærkninger til gentagelse af nonsens-stavelser:

- Endelsen er på ”u” eller ”a”
- Der forekommer ”v”, ”sj” og ”w”
- Lyt til listen og gentag
- Vi kan holde pause efter hver liste, men normalt kan man klare to til tre lister når man kommer ind i rytmen
- Forsøg at finde en behagelig sidestilling og bevæg dig derefter så lidt som muligt.

Bemærkninger til oplæsning af sætninger:

- Læs sætningen inden i dig selv
- Læs sætningen højt tre gange
- Forsøg at fortsætte selv om du laver fejl – det er næsten uundgåeligt

-

References

ANSI 1997. ANSI- S3.5-1997, Methods for the calculation of the speech Intelligibility index. American National Standards Institute, New York.

Henrichsen, P.J. (2009) "The CBS Text-to-Speech Workbench", CBS Working Papers on Linguistics Series, 26pp

Lane H, Tranel B., (1971). "The Lombard sign and the role of hearing in speech". J Speech Hear Res 14: 677–709. http://jslhr.asha.org/cgi/content/abstract/14/4/677.

Wells, J.C., (1997). 'SAMPA computer readable phonetic alphabet'. In Gibbon, D., Moore, R. and Winski, R. (eds.), 1997. *Handbook of Standards and Resources for Spoken Language Systems*. Berlin and New York: Mouton de Gruyter. Part IV, section B., www.phon.ucl.ac.uk/home/sampa